4

Systems
Optimization
Laboratory

A Modified Newton Method for Unconstrained Minimization

by
Anders L. Forsgren[*] Philip E. Gill[†] and Walter Murray[†]

TECHNICAL REPORT SOL 89-12

July 1989

Department of Operations Research
Stanford University
Stanford, CA 94305

4

SYSTEMS OPTIMIZATION LABORATORY
DEPARTMENT OF OPERATIONS RESEARCH
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4022

A Modified Newton Method for Unconstrained Minimization

by
Anders L. Forsgren[*], Philip E. Gill[†] and Walter Murray[†]

TECHNICAL REPORT SOL 89-12

July 1989

[*] Reseach supported by the Göran Gustafsson Foundation and the Swedish National Board for Technical Development.

[†] Research and reproduction of this report were partially supported by the National Science Foundation Grant ECS-8715153; U.S. Department of Energy Grant DE-FG03-87ER25030, and Office of Naval Research Contract N00014-87-K-0142.

Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do NOT necessarily reflect the views of the above sponsors.

Reproduction in whole or in part is permitted for any purposes of the United States Government. This document has been approved for public release and sale; its distribution is unlimited.

A MODIFIED NEWTON METHOD FOR UNCONSTRAINED MINIMIZATION

Anders L. FORSGREN[*] Philip E. GILL[†] and Walter MURRAY[‡]

[*]Optimization and Systems Theory, Department of Mathematics
The Royal Institute of Technology, S - 100 44 Stockholm, Sweden

[†]Department of Mathematics
University of California at San Diego, La Jolla, California 92093, USA

[‡]Systems Optimization Laboratory, Department of Operations Research
Stanford University, Stanford, California 94305-4022, USA

Technical Report SOL 89-12
July 1989

Abstract

Newton's method has proved to be a very efficient method for solving strictly convex unconstrained minimization problems. For the nonconvex case, various *modified* Newton methods have been proposed.

In this paper, a new modified Newton method is presented. The method is a linesearch method, utilizing the Cholesky factorization of a positive-definite portion of the Hessian matrix. The search direction is defined as a linear combination of a descent direction and a direction of negative curvature. Theoretical properties of the method are established and its behaviour is studied when applied to a set of test problems.

Keywords: Unconstrained minimization, modified Newton method, negative curvature, Cholesky factorization, linesearch, steplength algorithm

1. Introduction

In this paper we propose a method for finding a local minimizer of the problem

$$\underset{x\in\Re^n}{\text{minimize}} \quad f(x),$$

[*]Research supported by the Göran Gustafsson Foundation and the Swedish National Board for Technical Development.

[†‡] Research supported by the U.S. Department of Energy Grant DE-FG03-87ER25030; National Science Foundation Grant CCR-8413211; and the Office of Naval Research Contract N00014-87-K-0142.

where f is a twice-continuously differentiable function. This fundamental problem has been studied extensively and various methods have been proposed that use first and second derivatives. The aim is to generate a sequence of iterates $\{x_k\}_{k=0}^{\infty}$ that converge to a point $\bar{x}$ satisfying the first- and second-order necessary conditions, i.e., $\nabla f(\bar{x})$ is zero and $\nabla^2 f(\bar{x})$ is positive semidefinite.

Most methods that utilize second-derivative information may be viewed as extensions of Newton's method, in the sense that they are identical to Newton's method in a neighbourhood where the Hessian is positive definite. If the Hessian is not positive definite at some iterate, the Newton step may not reduce the objective function. Consequently, if the method is required to generate a sequence of improving estimates, some modification is needed. Such *modified* Newton methods have been studied for two decades, see for example Fiacco and McCormick [FM68], Gill and Murray [GM74], McCormick [McC77], Fletcher and Freeman [FF77], Mukai and Polak [MP78], Kaniel and Dax [KD79], Moré and Sorensen [MS79] and Goldfarb [Gol80].

Most modified Newton methods solve equations using a factorization of the Hessian. The method proposed by Gill and Murray [GM74] uses a modified Cholesky algorithm, in which a diagonal matrix is implicitly added to the Hessian to make it positive definite. A similar modified Cholesky algorithm based on an alternative diagonal correction has been proposed by Schnabel and Eskow [SE88]. The methods proposed by Fletcher and Freeman [FF77] and Moré and Sorensen [MS79] use the *Bunch-Parlett-Kaufman factorization* of the Hessian (see [BP71], [BK77]).

In the method proposed in this paper, the Cholesky algorithm with complete pivoting is performed until all potential pivot elements are smaller than a preassigned tolerance. The Cholesky factor is used to obtain a search direction, which may be a linear combination of a descent direction and a direction of negative curvature. It is shown that the gradient is zero and the smallest eigenvalue of the Hessian is bounded below by a small negative number at all limit points of the iterative sequence. The magnitude of the bound may be predetermined by adjusting certain preassigned tolerances.

2. Basics

2.1. Assumptions

The following assumptions are made throughout the paper:

A1. The objective function is twice continuously differentiable.

A2. The level set $S(x_0) = \{x : f(x) \leq f(x_0)\}$ associated with the starting point x_0 is compact.

2.2. Preassigned parameters

The proposed method depends on seven preassigned scalar parameters. These parameters specify different tolerances and for reference, their purpose and range of values are are briefly summarized here.

$\epsilon \in (0,1)$ is a parameter needed for the Cholesky factorization. It is used to determine the dimension of the positive-definite portion of the Hessian.

$h_{\min} > 0$ is a parameter used to reject small pivots in the factorization. No pivot elements smaller than $\epsilon^2 h_{\min}$ are accepted.

$\alpha_{\min}, \alpha_{\max}$ define an acceptable interval for the initial steplength.

$\eta \in (0,1)$ specifies a tolerance associated with the direction of negative curvature.

$\mu \in (0,\frac{1}{2})$ is a parameter used in the linesearch to guarantee a sufficient decrease in f.

$\gamma \in (0,1)$ is a parameter used to determine the rate of decrease of the steplength in the backtracking linesearch.

2.3. Terminology

The idea of a *descent direction* and a *direction of negative curvature* are important when computing the search direction. A vector p is a descent direction at a point x if $\nabla f(x)^T p < 0$. Likewise, p is a direction of negative curvature at x if $p^T \nabla^2 f(x) p < 0$.

Given a symmetric matrix

$$K = \begin{pmatrix} T & N^T \\ N & G \end{pmatrix},$$

with T nonsingular, the *Schur complement of* T *in* K will be denoted by K/T, and is defined as

$$K/T = G - NT^{-1}N^T.$$

The matrix K/T will be referred to as "the" Schur complement, when the matrix T is clear from the context. For further discussion of the Schur complement, see Cottle [Cot74].

Throughout the paper, the subscript k denotes the iteration index, and subscripts i and j denote particular components or columns of a matrix or vector. When element i, j of a matrix H_k is addressed, we refer to it as h_{ij}—i.e., the lowercase letter is used and the iteration subscript is dropped. Also, for vectors and matrices, when the term norm is used, we mean the Euclidean vector norm and the corresponding induced matrix norm.

3. Preliminary Discussion

At the k-th iteration of the proposed method, x_k denotes the current iteration point, g_k denotes $\nabla f(x_k)$ and H_k denotes $\nabla^2 f(x_k)$. With Newton's method as the model, it is desirable to compute the Newton search direction whenever H_k is sufficiently positive definite. If H_k is known to be positive definite, such a direction may be computed using the Cholesky factor of the Hessian. Whenever the Hessian is not sufficiently positive definite, the method presented here is based on the Cholesky factorization of a subset of the rows and columns of H_k. Complete pivoting is used, that is, the maximum diagonal element is chosen as the pivot at each step. Suppose that n_1 steps of the factorization have been performed and let Π denote the permutation matrix representing the column interchanges. We have

$$\Pi^T H_k \Pi = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} \quad \text{and} \quad \Pi^T g_k = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix}, \tag{3.1}$$

where H_{11} is a positive-definite prinicipal submatrix of order n_1, with Cholesky factor R_{11}. If $R_{12} = R_{11}^{-T} H_{12}$, we obtain the identity

$$\Pi^T H_k \Pi = \begin{pmatrix} R_{11}^T & 0 \\ R_{12}^T & I \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & \Pi^T H_k \Pi / H_{11} \end{pmatrix} \begin{pmatrix} R_{11} & R_{12} \\ 0 & I \end{pmatrix},$$

where $\Pi^T H_k \Pi / H_{11}$ is the Schur complement $H_{22} - H_{21} H_{11}^{-1} H_{12}$.

In order to simplify the notation, we shall assume that no permutations are required. This implies that Π is an identity matrix, and consequently $\Pi^T H_k \Pi = H_k$. We emphasize that this does not alter the theoretical results of later sections.

The factorization is usually terminated when all potential pivot elements in H_k/H_{11} are smaller than a tolerance $\epsilon^2 \max_i\{h_{ii}\}$. However, if all diagonal elements of the Hessian are small or negative, the pivot tolerance is given by $\epsilon^2 h_{\min}$ for a preassigned positive constant $h_{\min}$. Consequently, the pivot tolerance is defined as $\epsilon^2 h_k$, where

$$h_k = \max\{\max_i\{h_{ii}\}, h_{\min}\}. \tag{3.2}$$

The Cholesky factor is computed by rows, and the Schur complement is explicitly updated at each step of the factorization. Consequently, if the factorization is terminated with $n_1 < n$, the elements of the final Schur complement are known. Moreover, since we control the smallest acceptable pivot element, we have an upper bound on the diagonal elements of the $n_2 \times n_2$ matrix H_k/H_{11}. These properties of the factorization will prove important when computing directions of negative curvature. It is important to note that the dimensions of the matrices H_{11}, H_{12}, H_{21} and H_{22} depend on k.

The $n \times n_1$ matrix Z is defined to be

$$Z = \begin{pmatrix} I \\ 0 \end{pmatrix}, \tag{3.3}$$

where the matrix I is an $n_1 \times n_1$ identity matrix. The $n \times n_2$ matrix Y is defined to be

$$Y = \begin{pmatrix} -H_{11}^{-1} H_{12} \\ I \end{pmatrix}, \tag{3.4}$$

where we let y_j denote the j-th column of Y.

Lemma 3.1. *The following relations hold:*

$$\begin{aligned} Z^T H_k Z &= H_{11}, \\ Z^T H_k Y &= 0 \quad \textit{and} \\ Y^T H_k Y &= H_{22} - H_{21} H_{11}^{-1} H_{12}. \end{aligned}$$

Proof. The result follows from substituting for H_k/H_{11}, Z and Y using (3.1), (3.3) and (3.4). ∎

Again, we emphasize that the dimensions of Z and Y depend on k. The following lemma shows that the columns of the $n \times n$ matrix $M = \begin{pmatrix} Z & Y \end{pmatrix}$ form a basis for $\Re^n$.

Lemma 3.2. *The $n \times n$ matrix M is nonsingular.*

Proof. The result is immediate from the fact that $\det(M) = 1$. ∎

The following lemma relates the smallest eigenvalue of H_k to the smallest eigenvalue of H_k/H_{11}.

Lemma 3.3. *If H_k is indefinite then*

$$\lambda_{\min}(H_k/H_{11}) \leq \lambda_{\min}(H_k) \leq \frac{1}{\|Y\|^2}\lambda_{\min}(H_k/H_{11}).$$

Proof. It follows from Lemma 3.1 that $H_k/H_{11} = Y^T H_k Y$. Let u denote an eigenvector of unit length corresponding to the smallest eigenvalue of $Y^T H_k Y$. It follows from (3.4) that $u^T Y^T Y u \geq 1$. Sylvester's law of inertia yields $\lambda_{\min}(Y^T H_k Y) < 0$, giving

$$0 > \lambda_{\min}(Y^T H_k Y) = u^T Y^T H_k Y u = \frac{u^T Y^T H_k Y u}{u^T Y^T Y u} u^T Y^T Y u.$$

The proof of the second inequality is completed by noting that

$$0 > \frac{u^T Y^T H_k Y u}{u^T Y^T Y u} \geq \lambda_{\min}(H_k)$$

and

$$u^T Y^T Y u \leq \|Y\|^2.$$

Using the Courant-Fischer minimax characterization of eigenvalues (see e.g., Wilkinson [Wil65, page 101]), it follows that the smallest eigenvalue of H_k is the global minimum of the problem

$$\begin{array}{ll} \underset{v \in \Re^n}{\text{minimize}} & v^T H_k v \\ \text{subject to} & v^T v = 1. \end{array} \tag{3.5}$$

Lemma 3.2 implies the existence of vectors v_Z and v_Y such that $v = Z v_Z + Y v_Y$. Substituting for v in (3.5), and using the identity $Z^T H_k Y = 0$ yields the problem

$$\begin{array}{ll} \underset{v_Z \in \Re^{n_1}, v_Y \in \Re^{n_2}}{\text{minimize}} & v_Z^T Z^T H_k Z v_Z + v_Y^T Y^T H_k Y v_Y \\ \text{subject to} & v_Z^T Z^T Z v_Z + 2 v_Z^T Z^T Y v_Y + v_Y^T Y^T Y v_Y = 1. \end{array} \tag{3.6}$$

By definition, $Z^T H_k Z = H_{11}$ is positive definite, and it follows that the global minimum of (3.6) is no smaller than the global minimum of the problem

$$\begin{array}{ll} \underset{v_Z \in \Re^{n_1}, v_Y \in \Re^{n_2}}{\text{minimize}} & v_Y^T Y^T H_k Y v_Y \\ \text{subject to} & v_Z^T Z^T Z v_Z + 2 v_Z^T Z^T Y v_Y + v_Y^T Y^T Y v_Y = 1. \end{array} \tag{3.7}$$

Since this is a problem where the gradient of the constraint is nonzero at all feasible points, the constraint qualification always holds. Therefore, if v_Z and v_Y are global minimizers, there must exist a Lagrange multiplier ν such that the equations

$$\nu(Z^TZv_Z + Z^TYv_Y) = 0 \tag{3.8a}$$
$$Y^TH_kYv_Y + \nu(Y^TZv_Z + Y^TYv_Y) = 0 \tag{3.8b}$$
$$v_Z^TZ^TZv_Z + 2v_Z^TZ^TYv_Y + v_Y^TY^TYv_Y = 1 \tag{3.8c}$$

are satisfied.

The global minimum of (3.6) is negative, so that the global minimum of (3.7) is also negative. If ν is zero, it follows from (3.8b) that the global minimum of (3.7) is zero, which is a contradiction. Therefore, (3.8a) implies that v_Z is determined by v_Y, with

$$v_Z = -(Z^TZ)^{-1}Z^TYv_Y.$$

Using this value of v_Z and the definitions of Z and Y, problem (3.7) is equivalent to the problem

$$\begin{array}{ll} \underset{v_Y \in \Re^{n_2}}{\text{minimize}} & v_Y^TY^TH_kYv_Y \\ \text{subject to} & v_Y^Tv_Y = 1. \end{array} \tag{3.9}$$

The proof of the first inequality is completed by noting that the global minimum of (3.9) is the smallest eigenvalue of Y^TH_kY. ∎

We also require a result that relates the smallest eigenvalue of a symmetric matrix to the magnitude of its elements.

Lemma 3.4. *If all elements of an $n \times n$ symmetric matrix A have absolute values less than ρ, no eigenvalue of A has absolute value larger than ρn.*

Proof. This is an immediate consequence of the Gerschgorin circle theorem—see e.g., Golub and Van Loan [GV83, page 200]. ∎

4. The Cholesky Factorization

At each iterate, a positive-definite principal minor of the Hessian is factorized as outlined in the previous section. Some standard results concerning the Cholesky factorization are needed to derive uniform bounds on $\|H_{11}^{-1}\|$. These results are reviewed in this section. For a complete discussion of the Cholesky factorization, see Higham [Hig87].

Lemma 4.1. *If a positive-definite $n \times n$ matrix A is factorized using the Cholesky algorithm with complete pivoting, the elements of the Cholesky factor R have the following properties:*

$$r_{11} \geq r_{22} \geq \cdots \geq r_{nn}, \tag{4.1a}$$
$$|r_{ij}| < r_{ii} \quad \textit{for} \quad j = 1, \ldots, n, \quad i = 1, \ldots, j-1. \tag{4.1b}$$

Proof. For any $j > i$ the complete pivoting strategy yields

$$r_{jj}^2 \leq r_{ii}^2 - \sum_{l=i}^{j-1} r_{lj}^2,$$

from which (4.1a) follows. Since A is positive definite, it holds that r_{jj} is positive, and therefore $r_{ii}^2 > r_{ij}^2$. ∎

Lemma 4.2. *If R is the Cholesky factor of an $n \times n$ symmetric positive-definite matrix obtained by complete pivoting, the elements of its inverse U have the following properties.*

$$\begin{aligned} |u_{ij}| &< \frac{2^{j-i-1}}{r_{jj}} \quad && \textit{for} \quad j = 1, \ldots, n, \quad i = 1, \ldots, j-1 \\ u_{jj} &= \frac{1}{r_{jj}} && \textit{for} \quad j = 1, \ldots, n \\ u_{ij} &= 0 && \textit{for} \quad j = 1, \ldots, n, \quad i = j+1, \ldots, n. \end{aligned}$$

Proof. The matrix U satisfies the equation $RU = I$. The j-th column of this equation gives

$$\begin{aligned} u_{ij} &= 0 && \text{if} \quad i > j \\ u_{jj} &= \frac{1}{r_{jj}} \\ u_{ij} &= -\frac{1}{r_{ii}} \sum_{l=i+1}^{j} r_{il} u_{lj} && \text{if} \quad i < j. \end{aligned}$$

Lemma 4.1 implies that

$$|u_{ij}| < \sum_{l=i+1}^{j} |u_{lj}| \quad \text{if} \quad i < j.$$

By induction, it follows that

$$|u_{ij}| < \frac{2^{j-i-1}}{r_{jj}} \quad \text{if} \quad i < j.$$

∎

This bound on the element growth is usually unduly pessimistic. However, for certain special matrices, substantial element growth may occur—see e.g., Higham [Hig87, page 6]. What is important here is the *existence* of a bound. Such a bound is needed in order to obtain a uniform bound on $\|H_{11}^{-1}\|$.

Lemma 4.3. *There exists a positive constant c_0, such that for all k, $\|H_{11}^{-1}\| \leq c_0$.*

Proof. Since no pivot element smaller than $\epsilon^2 h_{\min}$ is accepted in the Cholesky factorization we have $r_{n_1 n_1} \geq \epsilon\sqrt{h_{\min}}$. Lemma 4.2 implies that

$$(R_{11}^{-1})_{ij} \leq \frac{2^n}{\epsilon\sqrt{h_{\min}}} \quad \text{for} \quad i = 1, \ldots, n_1 \quad \text{and} \quad j = 1, \ldots, n_1.$$

The identity $H_{11}^{-1} = R_{11}^{-1} R_{11}^{-T}$ and Lemma 3.4 yield the desired result. ∎

5. Computation of the Search Direction

In the proposed method a search direction p_k is computed at the k-th iterate. The vector p_k is defined in terms of two other vectors; a descent direction s_k and a direction of negative curvature d_k.

5.1. Computation of the descent direction

The descent direction s_k satisfies the equation

$$B_k s_k = -g_k, \tag{5.1}$$

where

$$B_k = \begin{pmatrix} H_{11} & 0 \\ 0 & h_k I \end{pmatrix}, \tag{5.2}$$

and h_k is defined by (3.2).

If $n_1 = n$, then $B_k = H_k$ and s_k is the Newton direction. If $n_2 = n$ then $B_k = h_{\min} I$ and s_k is a multiple of the steepest-descent direction. (In general, n_1 need not be equal to the number of positive eigenvalues of H_k. For example, the matrix $I - ee^T$, where e denotes an n-vector with unit components, has $n-1$ positive eigenvalues, but $n_1 = 0$. However, the results reported in Section 8 include only one case where n_1 was zero.)

The vector s_k of (5.1) is computed by solving the triangular systems

$$R_{11}^T u = -g_1 \quad \text{and} \quad R_{11} v = u, \quad \text{with} \quad s_k = \begin{pmatrix} v \\ -(1/h_k) g_2 \end{pmatrix}.$$

When s_k is computed from these equations, the norms of s_k and g_k are related in a uniform way. The following lemma shows that s_k satisfies descent properties similar to those required by McCormick [McC77].

Lemma 5.1. *If s_k is defined by (5.1) there exist positive constants c_1 and c_2 independent of k, such that for all k, it holds that*

$$-s_k^T g_k \geq c_1 \|g_k\|^2 \quad \textit{and} \quad \|g_k\| \geq c_2 \|s_k\|.$$

Proof. The definition of B_k yields

$$\|g_k\| \geq \frac{1}{\max\{\|H_{11}^{-1}\|, (1/h_k)\}}\|s_k\|.$$

The bound on $\|H_{11}^{-1}\|$ obtained from Lemma 4.3 implies the existence of c_2.

The definition of s_k yields

$$-s_k^T g_k \geq \min\{\lambda_{\min}(H_{11}^{-1}), (1/h_k)\}\|g_k\|^2.$$

Since H_{11} is positive definite and symmetric, we may employ the identity

$$\lambda_{\min}(H_{11}^{-1}) = \frac{1}{\lambda_{\max}(H_{11})} = \frac{1}{\|H_{11}\|}.$$

The compactness of $S(x_0)$ and the smoothness of f ensure the existence of c_1. ∎

5.2. Computation of the direction of negative curvature

The formula for d_k is derived from a method for computing directions of negative curvature in quadratic programming (see Forsgren *et al.* [FGM89]). If the variables corresponding to H_{22} are temporarily locked at their current values, a direction of negative curvature is defined by releasing one or two of the locked variables. This scheme corresponds to using either y_i or $y_i \pm y_j$ as a direction of negative curvature for a specific choice of i and j. The choice of i and j is determined by the values of the elements of H_k/H_{11}. When the factorization of H_k is terminated, these elements $y_i^T H_k y_j$ are known for $1 \leq i,\ j \leq n_2$, without explicitly computing the vectors y_i (see Lemma 3.1).

Let $\eta \in (0,1)$ denote a preassigned constant and let ρ denote $\max_{i,j} |y_i^T H_k y_j|$. The vector d_k is computed as follows.

if $\rho < \epsilon^2 h_k/\eta$ **then**

$$d_k = 0 \tag{5.3a}$$

else if $y_i^T H_k y_i = -\rho$ for some i **then**

$$d_k = \pm y_i \tag{5.3b}$$

else if $|y_i^T H_k y_j| = \rho$ for some $i \neq j$ **then**

$$d_k = \pm\frac{1}{\sqrt{2}}(y_i - \operatorname{sgn}(y_i^T H_k y_j) y_j) \tag{5.3c}$$

end if

In each case, we choose the sign of d_k so that $g_k^T d_k \leq 0$.

In the Cholesky algorithm, the pivot elements are chosen from the diagonal of the Schur complement, and it follows that $y_i^T H_k y_i \leq \epsilon^2 h_k$ for $i = 1, \ldots, n_2$. Consequently, if $\rho \geq \epsilon^2 h_k/\eta$ then $y_i^T H_k y_i < \rho$ for all i and d_k is well defined.

In order to obtain d_k, it is necessary to compute y_i or $y_i \pm y_j$. This is done by solving an equation involving R_{11} and R_{12}. For example, the computation of $y_i + y_j$ requires the solution of the equation

$$R_{11}u = -\frac{1}{\sqrt{2}}R_{12}(e_i + e_j) \quad \text{with} \quad y_i + y_j = \begin{pmatrix} u \\ \frac{1}{\sqrt{2}}(e_i + e_j) \end{pmatrix}. \tag{5.4}$$

The following lemma shows that any nonzero d_k is a direction of negative curvature.

Lemma 5.2. *If d_k is nonzero then*

$$g_k^T d_k \le 0 \quad \textit{and} \quad d_k^T H_k d_k \le -\frac{(1-\eta)\epsilon^2 h_k}{\eta}.$$

Proof. In each case, the sign of d_k is chosen so that $g_k^T d_k \le 0$. Let ρ denote $\max_{i,j} |y_i^T H_k y_j|$. If d_k is given by (5.3b), then $d_k^T H_k d_k = y_i^T H_k y_i = -\rho$. If d_k is given by (5.3c), then

$$d_k^T H_k d_k = \frac{1}{2}(y_i^T H_k y_i + y_j^T H_k y_j) - |y_i^T H_k y_j|,$$

where $|y_i^T H_k y_j| = \rho$. Since $y_i^T H_k y_i$ and $y_j^T H_k y_j$ are both less than or equal to $\epsilon^2 h_k$, it holds that $d_k^T H_k d_k \le \epsilon^2 h_k - \rho$. The inequality $\rho \ge \epsilon^2 h_k/\eta$ implies that in either case

$$d_k^T H_k d_k \le -\frac{(1-\eta)\epsilon^2 h_k}{\eta},$$

as required. ▮

Finally, we relate the curvature along any nonzero d_k to the smallest eigenvalue of H_k.

Lemma 5.3. *If d_k is nonzero, there exists a positive constant c_3, independent of k, such that for all k*

$$\frac{d_k^T H_k d_k}{d_k^T d_k} \le c_3 \, \lambda_{\min}(H_k).$$

Proof. Let ρ denote $\max_{i,j} |y_i^T H_k y_j|$. If d_k is nonzero, it follows from the proof of Lemma 5.2 that $d_k^T H_k d_k \le -(1-\eta)\rho$. Lemma 3.4 implies that $\lambda_{\min}(H_k/H_{11}) \ge -\rho n$, hence

$$d_k^T H_k d_k \le \frac{1-\eta}{n}\lambda_{\min}(H_k/H_{11}).$$

From Lemma 3.3 we have

$$\frac{d_k^T H_k d_k}{d_k^T d_k} \le \frac{1-\eta}{n \, d_k^T d_k}\lambda_{\min}(H_k).$$

Now (5.3b), (5.3c) and (5.4) may be used to obtain

$$1 \le d_k^T d_k \le \|Y^T Y\| \le 1 + \|H_{21}\|^2 \|H_{11}^{-1}\|^2.$$

The uniform boundedness of $\|d_k\|$ now follows from Lemma 4.3 and the assumptions on f. ∎

The significance of this lemma is that a nonzero d_k cannot be an arbitrarily poor direction of negative curvature compared to the eigenvector corresponding to the smallest eigenvalue of H_k (which is the best possible choice). The vector d_k may be zero even if H_k is indefinite. However, when d_k is zero, the following lemma gives a bound on the indefiniteness of H_k.

Lemma 5.4. *If $d_k = 0$ then $\lambda_{\min}(H_k) > -n\epsilon^2 h_k/\eta$.*

Proof. If $n_2 = 0$, then H_k is positive definite and $\lambda_{\min}(H_k) > 0$. Assume that $n_2 > 0$. Since $d_k = 0$, it holds that $\rho < \epsilon^2 h_k/\eta$. This result, together with Lemma 3.4 implies that

$$\lambda_{\min}(H_k/H_{11}) \geq -n_2\rho > -\frac{n\epsilon^2 h_k}{\eta},$$

and it follows from Lemma 3.3 that

$$\lambda_{\min}(H_k) > -\frac{n\epsilon^2 h_k}{\eta}.$$

∎

5.3. Computation of the search direction

The search direction p_k is defined to be

$$p_k = s_k + \beta_k d_k, \tag{5.5}$$

where the scalar β_k is defined as follows:

if $d_k \neq 0$ **and** $s_k^T H_k s_k \geq d_k^T H_k d_k$ **then**

$$\beta_k = -\frac{s_k^T H_k d_k}{d_k^T H_k d_k} + \sqrt{\left(\frac{s_k^T H_k d_k}{d_k^T H_k d_k}\right)^2 + 1 - \frac{s_k^T H_k s_k}{d_k^T H_k d_k}} \tag{5.6a}$$

else

$$\beta_k = 0 \tag{5.6b}$$

end if

Note that if $n_2 = 0$ (i.e., if H_k is sufficiently positive definite), p_k is the Newton direction. The choice of β_k is important only if d_k is nonzero. The particular choice of β_k given above is motivated by the following lemma, which also shows that if d_k is nonzero, p_k is also a direction of negative curvature.

Lemma 5.5. *If d_k is nonzero, then $\beta_k \geq 0$ and $p_k^T H_k p_k \leq d_k^T H_k d_k$.*

Proof. If $d_k \neq 0$ and $\beta_k = 0$ then $p_k^T H_k p_k = s_k^T H_k s_k \leq d_k^T H_k d_k$. If $d_k \neq 0$ and $\beta_k \neq 0$ it follows from the definition of β_k that $d_k^T H_k d_k \leq s_k^T H_k s_k$ and the square root in (5.6a) is well defined. In this case β_k is the unique positive number that satisfies the quadratic equation $(s_k + \beta_k d_k)^T H_k (s_k + \beta_k d_k) = d_k^T H_k d_k$. ∎

The following lemma shows that the norm of p_k is uniformly bounded.

Lemma 5.6. *If p_k is defined by (5.5), $\|p_k\|$ is uniformly bounded.*

Proof. Lemma 5.1 and the compactness of $S(x_0)$ imply that $\|s_k\|$ is uniformly bounded.

From the proof of Lemma 5.3 it follows that $\|d_k\|$ is uniformly bounded. Lemma 5.2 guarantees that the denominator of (5.6a) is uniformly bounded away from zero. Since $f \in C^2$ and the level set $S(x_0)$ is compact, it follows that β_k is uniformly bounded, as required. ∎

One consequence of Lemma 5.6 is that if d_k is nonzero, p_k cannot be an arbitrarily poor direction of negative curvature.

6. Computation of the Iterates

Unlike the methods suggested by McCormick [McC77], *Moré and Sorensen* [MS79] and Goldfarb [Gol80], if H_k is indefinite, the next iterate lies on a line emanating from x_k, instead of an arc. At a given iterate x_k, we will consider the case when an initial estimate $\alpha_k \in [\alpha_{\min}, \alpha_{\max}]$ of the steplength along p_k is given. One way of generating such an α_k is discussed in Section 8.1.

We follow McCormick [McC77] and guarantee a sufficient decrease by comparing f to a damped truncated Taylor series consisting of two or three terms. The resulting algorithm may be viewed as an Armijo-type linesearch [Arm66], extended to the indefinite case.

Let μ and γ denote preassigned constants such that $\mu \in (0, \frac{1}{2})$ and $\gamma \in (0, 1)$. Given x_k and $\alpha_k \in [\alpha_{\min}, \alpha_{\max}]$, the number i_k is defined to be the smallest nonnegative integer i such that

$$f(x_k + \gamma^i \alpha_k p_k) \leq f(x_k) + \mu \gamma^i \alpha_k g_k^T p_k \qquad \text{if} \quad d_k = 0; \quad (6.1a)$$

$$f(x_k + \gamma^i \alpha_k p_k) \leq f(x_k) + \mu \gamma^i \alpha_k g_k^T p_k + \frac{\mu^2 \gamma^{2i} \alpha_k^2}{2} p_k^T H_k p_k \quad \text{if} \quad d_k \neq 0. \quad (6.1b)$$

The next iterate x_{k+1} is defined as

$$x_{k+1} = x_k + \gamma^{i_k} \alpha_k p_k. \qquad (6.2)$$

A complete description of the modified Newton method is given in Algorithm 6.1. In order to show that the algorithm is well defined, we present two lemmas, which are slightly modified forms of a lemma given by Moré and Sorensen [MS79, Lemma 2.2].

```
Specify tolerances $\epsilon$, $h_{\min}$, $\alpha_{\min}$, $\alpha_{\max}$, $\eta$, $\mu$ and $\gamma$;
$k \leftarrow 0$;  converged ← false;
repeat
    Evaluate $f_k$, $g_k$ and $H_k$;
    Factorize $H_k$ to obtain $n_1$, $n_2$, $R_{11}$, $R_{12}$ and $H_k/H_{11}$;
    Compute $s_k$ and $d_k$;
    if ($n_1 = n$ or $d_k = 0$) then
        $p_k \leftarrow s_k$;
    else
        Compute $\beta_k$;
        $p_k \leftarrow s_k + \beta_k d_k$;
    end if
    converged ← convergence_test;
    if (not converged) then
        Compute $\alpha_k \in [\alpha_{\min}, \alpha_{\max}]$;
        Compute $i_k$ so that $f(x + \gamma^{i_k} \alpha_k p_k)$ is sufficiently decreased;
        $x_{k+1} \leftarrow x_k + \gamma^{i_k} \alpha_k p_k$;    $k \leftarrow k + 1$;
    end if
until converged;
```

Algorithm 6.1. *A modified Newton method for unconstrained minimization*

Lemma 6.1. *If $\mu \in (0, \frac{1}{2})$ is a given constant and φ is a continuously differentiable univariate function such that $\varphi'(0) < 0$, then there exists a positive scalar $\bar{\zeta}$ such that*

$$\varphi(\zeta) < \varphi(0) + \mu\varphi'(0)\zeta$$

for $\zeta \in (0, \bar{\zeta})$.

Proof. The Taylor-series expansion for a positive ζ yields

$$\frac{1}{\zeta}(\varphi(\zeta) - \varphi(0) - \mu\varphi'(0)\zeta) = (1 - \mu)\varphi'(0) + \varphi'(\theta\zeta) - \varphi'(0),$$

for some $\theta \in (0, 1)$, and it follows that

$$\lim_{\zeta \to 0+} \frac{1}{\zeta}(\varphi(\zeta) - \varphi(0) - \mu\varphi'(0)\zeta) = (1 - \mu)\varphi'(0) < 0.$$

Hence, there exists a positive number $\bar{\zeta}$ such that

$$\varphi(\zeta) - \varphi(0) - \mu\varphi'(0)\zeta < 0$$

for all $\zeta \in (0, \bar{\zeta})$. ∎

Lemma 6.2. *If $\mu \in (0, \frac{1}{2})$ is a given constant and φ is a twice-continuously differentiable univariate function such that $\varphi'(0) \le 0$ and $\varphi''(0) < 0$, then there exists a positive scalar $\bar{\zeta}$ such that*

$$\varphi(\zeta) < \varphi(0) + \mu\varphi'(0)\zeta + \mu^2\varphi''(0)\frac{\zeta^2}{2},$$

for $\zeta \in (0, \bar{\zeta})$.

Proof. The Taylor-series expansion for a positive ζ yields

$$\frac{1}{\zeta^2}(\varphi(\zeta) - \varphi(0) - \varphi'(0)\zeta - \mu^2\varphi''(0)\frac{\zeta^2}{2}) = \frac{1-\mu^2}{2}\varphi''(0) + \frac{1}{2}(\varphi''(\theta\zeta) - \varphi''(0))$$

for some $\theta \in (0,1)$, and it follows that

$$\lim_{\zeta \to 0+} \frac{1}{\zeta^2}(\varphi(\zeta) - \varphi(0) - \varphi'(0)\zeta - \mu^2\varphi''(0)\frac{\zeta^2}{2}) = \frac{1-\mu^2}{2}\varphi''(0) < 0.$$

Hence, there exists a positive number $\bar{\zeta}$ such that

$$\varphi(\zeta) - \varphi(0) - \varphi'(0)\zeta - \mu^2\varphi''(0)\frac{\zeta^2}{2} < 0,$$

for all $\zeta \in (0, \bar{\zeta})$. The proof is completed by noting that

$$\varphi'(0)\zeta \le \mu\varphi'(0)\zeta.$$

∎

We can now show that a sequence $\{x_k\}_{k=0}^{\infty}$ generated by (6.2) is well defined.

Lemma 6.3. *The sequence $\{x_k\}_{k=0}^{\infty}$ is well defined.*

Proof. First assume that d_k is nonzero. It follows from Lemmas 5.2 and 5.5 that $g_k^T p_k \le 0$ and $p_k^T H_k p_k < 0$. If we define $\varphi(\zeta) = f(x_k + \zeta p_k)$, we have $\varphi'(0) = g_k^T p_k$ and $\varphi''(0) = p_k^T H_k p_k$. Lemma 6.2 implies that given α_k, there exists a nonnegative integer i_k such that (6.1b) holds.

Assume that d_k is zero so that $p_k = s_k$. If $s_k = 0$, then (6.1a) holds for $i_k = 0$. If $s_k \neq 0$, then Lemma 5.1 implies that $g_k^T s_k < 0$. The application of Lemma 6.1 with $\varphi(\zeta) = f(x_k + \zeta p_k)$ implies that there exists an i_k such that (6.1a) holds. ∎

It is of interest to study the behaviour of $f(x)$ along p_k. It follows from Taylor-series expansion that

$$f(x_k + \zeta_k p_k) = f(x_k) + \zeta_k g_k^T p_k + \frac{\zeta_k^2}{2} p_k^T H_k p_k + r_k(x_k, p_k, \zeta_k),$$

where the remainder term is given by

$$r_k(x_k, p_k, \zeta_k) = \frac{\zeta_k^2}{2} p_k^T(\nabla^2 f(x_k + \theta_k\zeta_k p_k) - \nabla^2 f(x_k))p_k \qquad (6.3)$$

for some $\theta_k \in (0,1)$.

In the following lemma, we establish the behaviour of the remainder term as k tends to infinity and ζ_k tends to zero.

Lemma 6.4. *If* $\lim_{k\to\infty}\zeta_k = 0$ *then*

$$\lim_{k\to\infty}\frac{r_k(x_k,p_k,\zeta_k)}{\zeta_k^2} = 0.$$

Proof. Using properties of norms and (6.3) we get

$$\frac{|r_k(x_k,p_k,\zeta_k)|}{\zeta_k^2} \le \frac{\|p_k\|^2}{2}\|\nabla^2 f(x_k+\theta_k\zeta_k p_k) - \nabla^2 f(x_k)\|.$$

Assumption A2 and Lemma 5.6 imply that $\|x_k\|$ and $\|p_k\|$ are uniformly bounded. Since $\lim_{k\to\infty}\zeta_k = 0$ it follows that $|\zeta_k|$ is uniformly bounded. Therefore, there exists a compact set $\bar{C}$ such that $x_k \in \bar{C}$ and $x_k + \theta_k\zeta_k p_k \in \bar{C}$ for all k. Since $\bar{C}$ is compact and f is twice continuously differentiable, it follows that $\|\nabla^2 f\| : \bar{C} \to \Re$ is uniformly continuous. Hence, for all $\bar{\epsilon} > 0$ there exists a $\bar{\delta} > 0$ such that $\|\nabla^2 f(x) - \nabla^2 f(y)\| < \bar{\epsilon}$ for all $x, y \in \bar{C}$ such that $\|x - y\| < \bar{\delta}$. Since $\lim_{k\to\infty}\zeta_k = 0$ and $\|p_k\|$ is uniformly bounded, for each $\bar{\delta}$ there exists a $\bar{K}$ such that $\|\theta_k\zeta_k p_k\| < \bar{\delta}$ for all $k > \bar{K}$. ▮

If an infinite sequence $\{x_k\}_{k=0}^{\infty}$ is generated, the following lemma shows that there are only a finite number of iterates where a direction of negative curvature is computed.

Lemma 6.5. *For any sequence* $\{x_k\}_{k=0}^{\infty}$ *there must exist a finite* K *such that* $d_k = 0$ *for all* $k \ge K$.

Proof. The sequence $\{f(x_k)\}_{k=0}^{\infty}$ is decreasing and Assumptions A1 and A2 imply that this sequence is bounded from below, and it follows that $\{f(x_k)\}_{k=0}^{\infty}$ converges to a limit $\bar{f}$. Assume that there exists an infinite subsequence $\{x_k\}_{k\in J}$ such that $d_k \ne 0$ for all $k \in J$. From the equation

$$\bar{f} - f(x_0) = \sum_{k=0}^{\infty}(f(x_{k+1}) - f(x_k))$$

and the fact that each term in this sum is nonpositive, it follows that

$$\bar{f} - f(x_0) \le \sum_{k\in J}(f(x_{k+1}) - f(x_k)).$$

From Lemmas 5.2 and 5.5 we obtain

$$p_k^T H_k p_k \le d_k^T H_k d_k \le -\frac{(1-\eta)\epsilon^2 h_k}{\eta}$$

for all $k \in J$. The inequalities $g_k^T p_k \le 0$ and $\alpha_k \ge \alpha_{\min}$ imply that

$$\bar{f} - f(x_0) \le \sum_{k\in J} -\frac{\mu\gamma^{2i_k}\alpha_{\min}^2(1-\eta)\epsilon^2 h_{\min}}{2\eta}.$$

Since $\bar{f}$ is finite this inequality must imply that $i_k \to \infty$ as $k \to \infty$, for $k \in J$. Further, from the definition of i_k

$$f(x_k + \gamma^{i_k - 1}\alpha_k p_k) > f(x_k) + \mu\gamma^{i_k - 1}\alpha_k g_k^T p_k + \frac{\mu^2\gamma^{2(i_k-1)}\alpha_k^2}{2} p_k^T H_k p_k.$$

The Taylor-series expansion yields

$$\frac{r_k(x_k, p_k, \gamma^{i_k-1}\alpha_k)}{\gamma^{2(i_k-1)}\alpha_k^2} > -\frac{(1-\mu)}{\gamma^{i_k-1}\alpha_k} g_k^T p_k - \frac{(1-\mu^2)}{2} p_k^T H_k p_k.$$

Using the fact that $g_k^T p_k \leq 0$, it follows from Lemmas 5.2 and 5.5 that

$$\frac{r_k(x_k, p_k, \gamma^{i_k-1}\alpha_k)}{\gamma^{2(i_k-1)}\alpha_k^2} > \frac{(1-\mu^2)(1-\eta)\epsilon^2 h_{\min}}{2\eta}. \tag{6.4}$$

Taking the limit in (6.4) noting that $\alpha_k \leq \alpha_{\max}$ it follows from Lemma 6.4 that

$$0 \geq \frac{(1-\mu^2)(1-\eta)\epsilon^2 h_{\min}}{2\eta},$$

which is a contradiction. Therefore, there exists a finite K such that d_k is zero for all $k \geq K$. ∎

7. Global Convergence Properties

Using the established lemmas we can derive the following theorem concerning the limit points of the sequence $\{x_k\}_{k=0}^{\infty}$.

Theorem 7.1. *If an infinite sequence $\{x_k\}_{k=0}^{\infty}$ is generated as defined in (6.2), any limit point $\bar{x}$ satisfies*

$$\nabla f(\bar{x}) = 0 \quad \textit{and} \quad \lambda_{\min}(\nabla^2 f(\bar{x})) \geq -\frac{n\epsilon^2 \bar{h}}{\eta},$$

where $\bar{h} = \max\{\max_i\{(\nabla^2 f(\bar{x}))_{ii}\}, h_{\min}\}$

Proof. Without loss of generality, it may be assumed that the sequence $\{x_k\}_{k=0}^{\infty}$ converges to some point $\bar{x}$. Lemma 6.5 implies that there exists a K such that $d_k = 0$ for all $k \geq K$, so that $p_k = s_k$ for $k \geq K$. Therefore, Lemma 5.4 and the continuity of $\nabla^2 f$ imply that

$$\lambda_{\min}(\nabla^2 f(\bar{x})) \geq -\frac{n\epsilon^2 \bar{h}}{\eta}.$$

Assume that there exists an I such that $i_k < I$ for all $k \geq K$. It follows from Lemma 5.1 that

$$f(x_{k+1}) - f(x_k) \leq \mu\gamma^I \alpha_k g_k^T s_k \leq -\mu\gamma^I c_1 \alpha_{\min} \|g_k\|^2.$$

Since $f(\bar{x})$ is finite, it follows that $\nabla f(\bar{x}) = 0$.

If the integers i_k are not bounded above, then it may be assumed without loss of generality that $i_k \to \infty$ as $k \to \infty$. From the definition of i_k it follows that

$$f(x_k + \gamma^{i_k-1}\alpha_k s_k) - f(x_k) > \mu\gamma^{i_k-1}\alpha_k g_k^T s_k$$

for all $k \geq K$. The Taylor-series expansion yields

$$\frac{\gamma^{i_k-1}\alpha_k}{2} s_k^T H_k s_k + \frac{r_k(x_k, s_k, \gamma^{i_k-1}\alpha_k)}{\gamma^{i_k-1}\alpha_k} > -(1-\mu)g_k^T s_k.$$

Using Lemma 5.1 it follows that

$$\frac{\gamma^{i_k-1}\alpha_k}{2} s_k^T H_k s_k + \frac{r_k(x_k, s_k, \gamma^{i_k-1}\alpha_k)}{\gamma^{i_k-1}\alpha_k} > (1-\mu)c_1\|g_k\|^2.$$

Taking the limit and using Lemmas 5.6 and 6.4 we have $\nabla f(\bar{x}) = 0$ as required. ∎

As stated in the following corollary, a consequence of this theorem is that if two consequent iterates are identical, a limit point is found, since all subsequent iterates are identical.

Corollary 7.1. *If two consecutive iterates x_k and x_{k+1} are identical, the point x_k satisfies*

$$\nabla f(x_k) = 0 \quad \textit{and} \quad \lambda_{\min}(\nabla^2 f(x_k)) \geq -\frac{n\epsilon^2 h_k}{\eta}.$$

∎

The assumptions made are not sufficient to guarantee that the sequence $\{x_k\}_{k=0}^{\infty}$ is convergent. Some additional conditions are needed to guarantee that a generated sequence has a unique limit point. As observed by Moré and Sorensen [MS79], if we make the additional assumption that there are only a finite number of points in $S(x_0)$ where the gradient vanishes, the following result may be used.

Lemma 7.1. (Ortega and Rheinboldt [OR70]) *Suppose that a generated sequence $\{x_k\}_{k=0}^{\infty}$ satisfies*

$$\lim_{k\to\infty}(x_{k+1} - x_k) = 0 \quad \textit{and} \quad \lim_{k\to\infty}\nabla f(x_k) = 0.$$

Furthermore, suppose that the level set $S(x_0)$ is compact. If there are only a finite number of points in $S(x_0)$ where the gradient vanishes, then there exists a point $\bar{x}$ such that

$$\lim_{k\to\infty} x_k = \bar{x} \quad \textit{and} \quad \nabla f(\bar{x}) = 0.$$

Proof. See Ortega and Rheinboldt [OR70, Theorem 14.1.5]. ∎

In the method proposed in this paper, it follows from Lemma 6.5 that there is a K such that $p_k = s_k$ for $k \geq K$. From Lemma 5.1 we get $\lim_{k\to\infty} s_k = 0$. Using Lemma 7.1 the following corollary may be established.

Corollary 7.2. *If there are only a finite number of points in $S(x_0)$ where the gradient vanishes, the sequence $\{x_k\}_{k=0}^{\infty}$ converges to a point $\bar{x}$ satisfying*

$$\nabla f(\bar{x}) = 0 \quad \textit{and} \quad \lambda_{\min}(\nabla^2 f(\bar{x})) \geq -\frac{n\epsilon^2 \bar{h}}{\eta},$$

where $\bar{h} = \max\{\max_i\{(\nabla^2 f(\bar{x}))_{ii}\}, h_{\min}\}$. ∎

8. Test Problems and Numerical Results

A Fortran version of the algorithm was run on two types of test problems: nonlinear least-squares problems and barrier problems. The computer used was a DEC VAXstation II, with relative machine precision $\epsilon_M \approx 1.39 \times 10^{-17}$.

8.1. Parameter values

Various values of the parameters discussed in Section 2.2 were investigated. The results presented here were obtained with the following values:

ϵ	10^{-6}	(specifies smallest acceptable pivot element)
$h_{\min}$	10^{-3}	(smallest acceptable maximum diagonal element of H_{11})
η	10^{-3}	(tolerance for the acceptance of d_k)
$\alpha_{\min}$	10^{-10}	(minimum step in the linesearch)
$\alpha_{\max}$	10^{15}	(maximum step in the linesearch)
μ	0.1	(damping factor used in the truncated Taylor polynomial)
γ	0.5	(parameter for the backtracking).

The value of ϵ is a tradeoff between a small value that gives the Newton search direction when H_k is positive definite, and a value large enough to ensure that H_{11} is well-conditioned. Theoretically, a very small value of ϵ is preferred, since this is more likely to give limit points that satisfy the first- and second-order necessary conditions (see Theorem 7.1). However, small values of ϵ may give ill-conditioned Cholesky factors which may cause inaccurate search directions. Our experiments indicate that the overall performance of the method is not sensitive to the precise value of ϵ.

Given the value of ϵ, $h_{\min}$ is selected to ensure that the minimum pivot element is always greater than the machine precision. In the experiments presented here, this value of $h_{\min}$ affected only two iterates.

The value of η was varied by several orders of magnitude from the chosen value, without changing the overall performance. The value selected helps to avoid computing directions of negative curvature when the elements of the Schur complement are all small in magnitude.

The steplength α_k is computed using the linesearch procedure of Gill *et al.* [GMSW79] with default parameter settings. At each step of the linesearch both the function and gradient are evaluated. The value of μ above was chosen to ensure that $i_k = 0$ is accepted in most cases. Since the value of i_k in (6.2) differed from zero in only two cases, we deduce that the choice of γ is not crucial. The values of $\alpha_{\min}$ and $\alpha_{\max}$ were designed to ensure that the steplength produced by the linesearch is accepted in almost all cases. In practice, a sensible choice of $\alpha_{\max}$ can improve efficiency.

The efficiency of the linesearch is affected by the initial estimate of α. Whenever d_k was zero, the choice of $\alpha = 1$ was found to be adequate. However, the unit step tended to overestimate the accepted step when d_k was nonzero. To allow for this, an initial step of 0.01 was used in these cases.

8.2. Least-squares test problems

The least-squares test problems comprise a suite of 45 problems, given by Fraley [Fra88]. Many of these problems are known to be hard to solve, in spite of their small size. A summary of results obtained on these problems when applying different least-squares methods and methods for unconstrained minimization is given in Fraley [Fra88]. Our numbering of the problems is the same as in Fraley's study. The formulations for problems 1–35*c* are given by Moré *et al.* [MGH81], problems 36*a*–36*d* are presented in Fraley [Fra88], problems 37–38 are given by Salane [Sal87], problems 39*a*–41*g* are from McKeown [McK75], problems 42*a*–43*f* originate from de Villiers and Glasser [dVG81] and problems 44*a*–45*e* are from Dennis *et al.* [DGV85].

We accept x_k as a solution of a least-squares problem if one of the following two conditions are met:

C1.
$$\begin{aligned} d_k &= 0 \\ \|g_k\| &\le \sqrt{\epsilon_M} \end{aligned}$$

or

C2.
$$\begin{aligned} d_k &= 0 \\ f(x_{k-1}) - f(x_k) &\le \epsilon_M(1 + |f(x_k)|) \\ \|x_k - x_{k-1}\| &\le \sqrt{\epsilon_M}(1 + \|x_k\|) \\ \|g_k\| &\le \sqrt[3]{\epsilon_M}(1 + |f(x_k)|). \end{aligned}$$

The first condition is intended to accept points that approximately satisfy the first- and second-order necessary conditions for optimality. The second condition is intended to test when the sequence $\{x_k\}_{k=0}^{\infty}$ has converged. For a detailed discussion of convergence criteria for unconstrained optimization, see Gill *et al.* [GMW81, Chapter 8].

In some problems it was not possible to evaluate the function at all tiial points. In these cases, the trial step was repeatedly decreased by a factor γ ($\gamma = 0.5$) until f could be evaluated. This additional backtracking was necessary for problems 42*a* and 43*d* because of an implicit nonnegativity constraint on one variable; and for problem 19 because of overflow during the calculation of the objective function. Similarly, the initial step at the starting point of problem 11 was repeatedly decreased until the Hessian and gradient were not numerically zero. These trial function evaluations are included in the number of function evaluations shown.

In problems 2, 36*a*, 36*b* and 36*d*, the algorithm failed to converge within the permitted number of iterations. In all cases, this non-convergence is a consequence of the Hessian being very ill-conditioned at the solution. Although the algorithm did not converge in these cases, the objective value was close to the optimal objective value.

nr	*name*	n	n_1	f_k	$\|g_k\|$	$\kappa(H_{11})$	k	*nf*	*conv*	$\#n_2^+$	$\#d_u$
1	rose	2	2	6.317050E-32	6.0E-15	2.E+03	22	29	y 0	0	0
2	froth	2	2	2.449213E+01	1.7E-07	1.E+03	7	11	y 2	0	0
3	powlbs	2	1	2.837351E-05	1.1E+01	1.E+00	600	2079	n 5	593	1
4	brownbs	2	1	3.851860E-34	2.8E-11	1.E+00	4	5	y 1	3	0
5	beale	2	2	1.007290E-23	3.8E-12	1.E+02	8	18	y 0	3	3
6	jensam	2	2	6.218109E+01	2.0E-13	8.E+00	10	11	y 0	0	0
7	helix	3	3	2.943716E-35	2.5E-17	3.E+02	14	25	y 0	5	5
8	bard	3	3	4.107439E-03	4.4E-16	2.E+03	14	21	y 0	1	1
9	gauss	3	3	5.639664E-09	4.9E-11	5.E+01	2	3	y 0	0	0
10	meyer	3	2	2.661418E+04	4.1E+01	6.E+07	24	74	n 4	23	2
11	gulf	3	3	8.612303E-20	2.0E-10	1.E+10	151	251	y 0	8	7
12	box	3	3	8.939108E-30	1.3E-15	7.E+03	14	19	y 0	1	0
13	sing	4	4	1.300559E-13	1.8E-09	1.E+08	21	22	y 0	0	0
14	wood	4	4	0.000000E+00	0.0E+00	5.E+02	39	52	y 0	1	1
15	kowosb	4	4	1.537528E-04	3.6E-11	2.E+03	9	23	y 0	4	4
16	brownden	4	4	4.291110E+04	1.6E-10	6.E+01	8	9	y 0	0	0
17	osb1	5	5	2.732447E-05	3.6E-09	1.E+09	65	147	y 0	28	28
18	exp6	6	5	2.827825E-03	1.9E-09	1.E+05	48	136	y 1	46	37
19	osb2	11	11	2.006887E-02	2.2E-12	4.E+03	16	37	y 0	6	6
20a	watson06	6	6	1.143835E-03	5.2E-13	2.E+04	12	13	y 0	0	0
20b	watson09	9	9	6.998801E-07	7.5E-15	2.E+08	13	14	y 0	0	0
20c	watson12	12	11	4.178499E-09	6.4E-08	8.E+09	31	38	y 3	32	1
20d	watson20	20	13	6.886510E-08	1.8E-08	2.E+11	53	107	y 3	54	0
21a	rosex	10	10	3.158525E-31	1.3E-14	2.E+03	22	29	y 0	0	0
21b	rosex2	20	20	6.317050E-31	1.9E-14	2.E+03	22	29	y 0	0	0
22a	singx	12	12	3.901678E-13	3.2E-09	1.E+08	21	22	y 0	0	0
22b	singx2	20	20	1.284503E-13	1.2E-09	2.E+08	22	23	y 0	0	0
23a	peni4	4	4	1.124989E-05	7.5E-11	5.E+03	34	43	y 0	0	0
23b	peni10	10	10	3.543826E-05	1.3E-12	1.E+03	36	44	y 0	0	0
24a	penii4	4	4	4.688147E-06	1.1E-10	2.E+06	110	158	y 0	0	0
24b	penii10	10	10	1.468303E-04	1.0E-09	2.E+06	93	132	y 0	0	0
25a	vardim1	10	10	8.680345E-27	2.6E-12	1.E+02	14	15	y 0	0	0
25b	vardim2	20	20	0.000000E+00	0.0E+00	4.E+02	18	19	y 0	0	0
26a	trig	10	10	1.721941E-24	1.3E-12	8.E+00	7	11	y 0	1	1
26b	trig2	20	20	3.074585E-28	1.2E-14	4.E+00	11	22	y 0	5	5
27a	brownal1	10	10	2.651544E-28	2.3E-14	2.E+03	8	9	y 0	0	0
27b	brownal2	20	20	2.462302E-18	3.3E-09	1.E+04	9	10	y 0	0	0
28a	discbv1	10	10	9.287387E-25	1.7E-13	9.E+01	3	4	y 0	0	0
28b	discbv2	20	20	1.182787E-25	1.7E-14	7.E+02	3	4	y 0	0	0
29a	discie1	10	10	1.997048E-22	2.5E-11	1.E+00	3	4	y 0	0	0
29b	discie2	20	20	3.293268E-22	3.2E-11	1.E+00	3	4	y 0	0	0
30a	broytri1	10	10	8.955574E-33	7.6E-16	2.E+00	6	7	y 0	0	0
30b	broytri2	20	20	2.051115E-32	1.1E-15	2.E+00	6	7	y 0	0	0
31a	broyban1	10	10	6.032100E-27	5.2E-13	3.E+00	8	9	y 0	0	0
31b	broyban2	20	20	6.067580E-27	5.2E-13	3.E+00	8	9	y 0	0	0
32	lin	10	10	5.000000E+00	9.7E-16	1.E+00	1	2	y 0	0	0
33	lin1	10	1	2.317073E+00	2.8E-11	1.E+00	1	3	y 1	2	0
34	lin0	10	1	3.067568E+00	4.0E-11	1.E+00	1	3	y 1	2	0
35a	chebyqu1	8	8	1.758437E-03	5.4E-15	2.E+01	19	34	y 0	11	11

Table 8.1: Results for least-squares test problems 1–35a.

nr	$name$	n	n_1	f_k	$\|g_k\|$	$\kappa(H_{11})$	k	nf	$conv$	$\#n_2^+$	$\#d_u$
35b	chebyqu2	9	9	9.668790E-22	8.5E-11	2.E+02	34	84	y 0	27	27
35c	chebyqu3	10	10	3.251977E-03	6.4E-11	2.E+02	24	46	y 0	16	16
36a	msqrt1i	4	4	7.839519E-11	5.5E-06	6.E+10	600	885	n 5	0	0
36b	msqrt2i	9	7	2.066297E-09	1.8E-05	2.E+08	600	2175	n 5	421	361
36c	msqrt3i	9	8	6.499547E-16	2.1E-08	4.E+08	28	44	n 4	2	1
36d	msqrt4i	9	9	2.105165E-09	1.6E-06	5.E+10	600	2154	n 5	415	361
37	han1	2	2	1.043501E+02	1.5E-12	3.E+04	5	9	y 0	1	0
38	han2	3	3	1.983216E+01	2.7E-10	1.E+06	6	11	y 0	0	0
39a	mck1a	2	2	9.180060E-02	1.0E-17	7.E+00	3	4	y 0	0	0
39b	mck1b	2	2	9.180060E-02	7.7E-12	5.E+00	3	4	y 0	0	0
39c	mck1c	2	2	9.180060E-02	1.3E-13	2.E+00	3	4	y 0	0	0
39d	mck1d	2	2	9.180060E-02	2.2E-17	2.E+00	5	6	y 0	0	0
39e	mck1e	2	2	9.180060E-02	1.0E-14	6.E+00	7	8	y 0	0	0
39f	mck1f	2	2	9.180060E-02	3.9E-18	7.E+00	10	11	y 0	0	0
39g	mck1g	2	2	9.180060E-02	8.8E-10	7.E+00	12	13	y 0	0	0
40a	mck2a	3	3	3.982776E-01	4.2E-15	2.E+01	3	4	y 0	0	0
40b	mck2b	3	3	3.982776E-01	1.2E-10	8.E+00	3	4	y 0	0	0
40c	mck2c	3	3	3.982776E-01	6.6E-13	2.E+00	4	5	y 0	0	0
40d	mck2d	3	3	3.982776E-01	1.0E-16	4.E+00	5	6	y 0	0	0
40e	mck2e	3	3	3.982776E-01	6.7E-17	1.E+01	7	8	y 0	0	0
40f	mck2f	3	3	3.982776E-01	5.4E-12	2.E+01	9	10	y 0	0	0
40g	mck2g	3	3	3.982776E-01	1.4E-15	2.E+01	12	13	y 0	0	0
41a	mck3a	5	5	5.000001E-01	8.3E-10	4.E+00	2	3	y 0	0	0
41b	mck3b	5	5	5.000001E-01	6.7E-14	3.E+00	3	4	y 0	0	0
41c	mck3c	5	5	5.000001E-01	4.8E-12	3.E+00	7	8	y 0	0	0
41d	mck3d	5	5	5.000001E-01	9.6E-15	2.E+00	8	9	y 0	0	0
41e	mck3e	5	5	5.000001E-01	3.7E-10	2.E+00	10	11	y 0	0	0
41f	mck3f	5	5	5.000001E-01	1.7E-11	3.E+00	13	14	y 0	0	0
41g	mck3g	5	5	5.000001E-01	2.1E-12	3.E+00	16	17	y 0	0	0
42a	devg1a	4	4	3.593754E-28	7.3E-12	5.E+04	16	27	y 0	2	2
42b	devg1b	4	4	2.485558E-23	1.2E-09	5.E+04	28	51	y 0	6	6
42c	devg1c	4	4	2.223602E-28	5.3E-12	5.E+04	21	43	y 0	5	5
42d	devg1d	4	4	1.910276E-28	7.0E-12	5.E+04	19	26	y 0	2	2
43a	devg2a	5	5	1.390367E-29	1.4E-12	8.E+06	17	26	y 0	3	3
43b	devg2b	5	5	1.352306E-25	9.4E-11	8.E+06	16	29	y 0	4	4
43c	devg2c	5	5	5.445605E-29	5.4E-12	8.E+06	13	27	y 0	6	5
43d	devg2d	5	5	9.207747E-22	2.1E-09	8.E+06	29	50	y 0	5	4
43e	devg2e	5	5	1.059680E-21	2.8E-09	8.E+06	17	30	y 0	5	5
43f	devg2f	5	5	3.254051E-30	2.8E-13	8.E+06	18	32	y 0	4	4
44a	dgv6a	6	6	3.982829E-24	2.3E-11	7.E+06	38	121	y 0	29	29
44b	dgv6b	6	6	1.255706E-31	1.6E-14	4.E+02	12	26	y 0	4	4
44c	dgv6c	6	6	8.742151E-25	8.2E-10	4.E+11	392	833	y 0	385	83
44d	dgv6d	6	6	3.416587E-26	1.2E-10	2.E+10	316	764	y 0	306	126
44e	dgv6e	6	6	1.306575E-30	1.1E-12	1.E+08	175	516	y 0	164	163
45a	dgv8a	8	8	5.542109E-26	1.4E-11	7.E+06	39	116	y 0	31	31
45b	dgv8b	8	8	3.710801E-33	6.7E-16	2.E+03	16	36	y 0	7	7
45c	dgv8c	8	8	1.234906E-30	1.3E-11	9.E+11	484	1003	y 0	480	134
45d	dgv8d	8	8	1.970398E-30	3.4E-12	4.E+10	480	1053	y 0	470	174
45e	dgv8e	8	8	3.968786E-31	6.0E-13	4.E+08	349	953	y 0	339	338

Table 8.2: Results for least-squares test problems 35b–45e.

nr	*name*	10^{-1}	10^{-2}	10^{-3}	10^{-4}
1	rose	5	12	16	17
2	froth	2	3	4	4
3	powlbs	1	4	8	235
4	brownbs	3	3	3	3
5	beale	2	3	4	5
6	jensam	2	3	5	6
7	helix	4	6	7	8
8	bard	2	4	6	8
9	gauss	1	1	1	1
10	meyer	1	2	2	3
11	gulf	1	1	3	6
12	box	2	3	5	7
13	sing	2	3	5	6
14	wood	1	3	4	26
15	kowosb	3	5	6	7
16	brownden	2	4	5	5
17	osb1	12	28	32	37
18	exp6	10	25	27	32
19	osb2	7	9	11	12
20a	watson06	1	2	5	7
20b	watson09	1	2	5	8
20c	watson12	1	3	4	6
20d	watson20	4	7	12	17
21a	rosex	5	12	16	17
21b	rosex2	5	12	16	17
22a	singx	2	3	5	6
22b	singx2	2	3	5	6
23a	peni4	2	3	5	6
23b	peni10	2	3	5	6
24a	penii4	2	2	3	4
24b	penii10	2	3	4	5
25a	vardim1	2	3	5	6
25b	vardim2	2	3	5	6
26a	trig	3	4	4	5
26b	trig2	6	7	8	8
27a	brownal1	1	1	1	1
27b	brownal2	1	1	1	1
28a	discbv1	1	1	1	2
28b	discbv2	1	1	1	2
29a	discie1	1	1	2	2
29b	discie2	1	1	2	2
30a	broytri1	1	2	3	3
30b	broytri2	1	2	3	3
31a	broyban1	2	3	4	5
31b	broyban2	2	3	4	5
32	lin	1	1	1	1
33	lini	1	1	1	1
34	lin0	1	1	1	1
35a	chebyqu1	10	15	15	16
35b	chebyqu2	20	29	29	31
35c	chebyqu3	11	14	19	21
36a	msqrt1i	2	4	6	10
36b	msqrt2i	2	4	6	10
36c	msqrt3i	2	3	5	7
36d	msqrt4i	2	4	6	10
37	han1	1	2	2	3
38	han2	1	2	3	3
39a	mck1a	1	1	1	1
39b	mck1b	1	1	1	2
39c	mck1c	1	1	1	2
39d	mck1d	2	2	3	3
39e	mck1e	2	3	4	5
39f	mck1f	2	3	5	6
39g	mck1g	2	3	5	6
40a	mck2a	1	1	1	1
40b	mck2b	1	1	2	2
40c	mck2c	1	2	2	2
40d	mck2d	1	2	2	3
40e	mck2e	2	3	4	4
40f	mck2f	2	3	5	6
40g	mck2g	2	3	5	6
41a	mck3a	1	1	1	1
41b	mck3b	1	1	1	1
41c	mck3c	2	3	4	5
41d	mck3d	1	3	4	5
41e	mck3e	2	4	5	6
41f	mck3f	2	3	5	6
41g	mck3g	2	3	5	6
42a	devg1a	8	10	12	12
42b	devg1b	20	23	24	25
42c	devg1c	14	16	17	18
42d	devg1d	11	14	14	15
43a	devg2a	2	3	5	6
43b	devg2b	3	5	7	9
43c	devg2c	1	4	7	7
43d	devg2d	4	6	8	10
43e	devg2e	2	4	7	9
43f	devg2f	3	6	8	10
44a	dgv6a	11	22	30	33
44b	dgv6b	3	6	8	9
44c	dgv6c	1	5	29	119
44d	dgv6d	1	3	24	97
44e	dgv6e	1	3	14	52
45a	dgv8a	11	22	29	33
45b	dgv8b	2	5	9	12
45c	dgv8c	6	9	27	99
45d	dgv8d	3	6	20	89
45e	dgv8e	3	4	14	57

Table 8.3: Number of iterations required to reduce $(f(x_k) - f(x^*))/(f(x_0) - f(x^*))$ below four different tolerances.

Ill-conditioning was also responsible for the failure in problems 10 and 36*c*. In these cases, the algorithm terminated because of a failure in the linesearch. Again, the objective value has been reduced significantly. In problem 36c the algorithm terminated at a point very close to the solution. In problem 10 the Hessian at the final iterate is positive definite but very ill-conditioned.

The results of the computer runs are summarized in Tables 8.1 and 8.2. The column headings have the following meaning:

nr	Problem number.
name	Problem name.
n	Number of variables.
n_1	Dimension of H_{11} at the final iterate x_k.
f_k	Value of the objective function the final iterate x_k.
$\|g_k\|$	Norm of the gradient at the final iterate x_k.
$\kappa(H_{11})$	Estimate of the condition number of the final H_{11}.
k	Number of iterations.
nf	Number of function evaluations.
conv	Convergence information.
	y 0 Convergence criteria C1 satisfied with $n_2 = 0$.
	y 1 Convergence criteria C1 satisfied with $n_2 > 0$.
	y 2 Convergence criteria C2 satisfied with $n_2 = 0$.
	y 3 Convergence criteria C2 satisfied with $n_2 > 0$.
	n 4 Nonconvergent due to failure in linesearch.
	n 5 Nonconvergent due to too many iterations (> 600).
$\#n_2^+$	Number of iterates where n_2 was positive.
$\#d_u$	Number of iterates where d was used.

Our experience from working on these problems is that it is possible to reduce the value of the objective function significantly in a relatively small number of iterations, as illustrated in Table 8.3. However, stringent convergence criteria such as those used here may not always be achievable if the Hessian is ill-conditioned at the solution.

8.3. Barrier test problems

The test problems with a general objective form originate from the barrier function approach of Resende *et al.* [RKR89] for solving 0–1 integer programming problems. The aim of this approach is to find a point x^* with all components ±1 in the set F, where F is defined to be

$$F = \left\{ x : \begin{pmatrix} A \\ -I \\ I \end{pmatrix} x \le \begin{pmatrix} 2b - Ae + e \\ e \\ e \end{pmatrix} \right\}, \qquad (8.1)$$

for an $m \times n$ matrix A and an n-vector b. We consider the case where all elements of A and b are integers. The vector e denotes a suitably dimensioned vector with unit components.

If the composite matrix and vector associated with the inequalities of (8.1) are denoted by $\bar{A}$ and $\bar{b}$, we may write $F = \{x : \bar{A}x \le \bar{b}\}$.

This integer feasibility problem is converted into a smooth minimization problem. The function to be minimized is the barrier function f defined by

$$f(x) = \frac{1}{2}\ln(n - x^T x) - \frac{1}{m+2n}\sum_{i=1}^{m+2n} \ln(e_i^T(\bar{b} - \bar{A}x))$$

(see Resende *et al.* [RKR89]). This barrier function does not satisfy the assumptions of Section 2.1, since the function is only defined for x such that $\bar{A}x < \bar{b}$. Moreover, as is shown in the appendix, the barrier function tends to minus infinity for a sequence converging to a point with all components ±1. Nevertheless, these functions are useful as test problems because they have many local minimizers and exhibit many directions of negative curvature. (Moreover, it was also of interest to see if the algorithm was able to locate a point in F with all components ±1.)

Three different test problems were used, and for each of them the set of points in F with all components ±1 consists of only one point, x^*.

Data for barrier test problem 1:

$$A = \begin{pmatrix} -2 & -1 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & -2 & -1 & 0 \\ 0 & -1 & 0 & -1 & 0 & -1 \\ 0 & 0 & -2 & 0 & -1 & -1 \\ 3 & 2 & 3 & 4 & 2 & 3 \end{pmatrix}, \quad b = \begin{pmatrix} -1 \\ -2 \\ -2 \\ -1 \\ 8 \end{pmatrix},$$

$$\begin{aligned} \text{a)} \quad x_0 &= (\ -0.90 \quad 0.76 \quad -0.76 \quad 0.64 \quad 0.20 \quad -0.20\ )^T, \\ \text{b)} \quad x_0 &= (\ -0.86 \quad 0.64 \quad -0.64 \quad 0.46 \quad -0.20 \quad 0.20\ )^T, \\ x^* &= (\ -1 \quad 1 \quad -1 \quad 1 \quad 1 \quad -1\ )^T. \end{aligned}$$

Data for barrier test problem 2:

$$A = \begin{pmatrix} 1 & 2 & 4 & 3 \\ -4 & -3 & -4 & -2 \end{pmatrix}, \quad b = \begin{pmatrix} 5 \\ -8 \end{pmatrix},$$

$$\begin{aligned} \text{a)} \quad x_0 &= (\ 0.90 \quad -0.10 \quad 0.45 \quad -0.95\ )^T, \\ \text{b)} \quad x_0 &= (\ 0.88 \quad 0.08 \quad 0.34 \quad -0.94\ )^T, \\ x^* &= (\ 1 \quad -1 \quad 1 \quad -1\ )^T. \end{aligned}$$

Data for barrier test problem 3:

$$A = \begin{pmatrix} 4 & 8 & 2 & 4 \\ 2 & 4 & 4 & 8 \\ -4 & -8 & -1 & -2 \end{pmatrix}, \quad b = \begin{pmatrix} 11 \\ 13 \\ -9 \end{pmatrix},$$

$$\begin{aligned} \text{a)} \quad x_0 &= (\ -0.40 \quad 0.80 \quad 0.20 \quad -0.99\ )^T, \\ \text{b)} \quad x_0 &= (\ -0.34 \quad 0.78 \quad 0.12 \quad -0.99\ )^T, \\ x^* &= (\ -1 \quad 1 \quad 1 \quad -1\ )^T. \end{aligned}$$

nr	*name*	*n*	n_I	f_k	$\lVert g_k \rVert$	$\kappa(H_{11})$	*k*	*nf*	*conv*	$\#n_2^+$	$\#d_u$
46a	barlog1a	6	5	-1.841628E+00	1.3E+07	3.E+00	18	22	y 6	19	19
46b	barlog1b	6	6	7.626996E-01	2.8E-12	3.E+01	7	11	y 0	3	3
47a	barlog2a	4	3	-1.122621E+00	9.0E+06	2.E+00	16	17	y 6	17	17
47b	barlog2b	4	4	5.805715E-01	3.5E-14	1.E+01	7	8	y 0	1	1
48a	barlog3a	4	3	-1.996615E+00	6.0E+06	3.E+00	16	17	y 6	17	17
48b	barlog3b	4	4	1.433882E-01	8.9E-15	8.E+00	10	14	y 0	2	2
49a	bar1a	6	5	1.618634E-01	2.3E+06	3.E+00	17	21	y 6	18	18
49b	bar1b	6	6	2.144056E+00	7.0E-12	3.E+01	7	11	y 0	3	3
50a	bar2a	4	3	3.771148E-01	8.2E+05	2.E+00	14	15	y 6	15	15
50b	bar2b	4	4	1.787059E+00	4.8E-13	1.E+01	7	8	y 0	1	1
51a	bar3a	4	3	1.435501E-01	1.1E+06	2.E+00	15	16	y 6	16	16
51b	bar3b	4	4	1.154178E+00	1.1E-11	8.E+00	10	13	y 0	3	3

Table 8.4: Results for barrier test problems

For each of the three test problems, two starting points were used. Problems $46a, 47a$ and $48a$ correspond to starting points for which the sequence $\{x_k\}_{k=0}^{\infty}$ converged to x^*. Problems $46b, 47b$ and $48b$ correspond to starting points for which the sequence converged to a local minimizer of the barrier function.

Problems 49, 50 and 51 are similar to problems 46, 47 and 48, except that the objective function is the argum' nt of the logarithmic barrier function, i.e.,

$$f(x) = \frac{(n - x^T x)^{1/2}}{(\prod_{i=1}^{m+2n} e_i^T(\bar{b} - \bar{A}x))^{1/(m+2n)}}.$$

For the same starting points, the same final points were reached in approximately the same number of iterations.

The barrier problems are not truly unconstrained, since the objective function is only defined for x such that $\bar{A}x < \bar{b}$. To allow for this, the iteration was modified so that $\alpha_{\max}$ was made subject to being no greater than 99.99% of the step to the boundary of F. If d_k was zero, the unit initial steplength was chosen, and when d_k was nonzero, a trial value of $0.8\,\alpha_{\max}$ was used. The trial step was accepted if the directional derivative was still negative. Otherwise, the same linesearch as used for the least-squares test problems was used.

The following criterion was used to decide when a point in F with all components ± 1 had been reached,

C3. $\max_i \left\{1 - |e_i^T x_k|\right\} \le 10\sqrt{\epsilon_M}$.

The results for the barrier test problems are presented in Table 8.4. The column headings are the same as for the least-squares test problem, and the only difference is that convergence criteria C3 is denoted by "y 6" in the *conv* column.

8.4. Practical behaviour of the computed directions

In Section 5 of this paper theoretical properties of the computed directions s_k, d_k and p_k are established. It is shown in Lemma 5.3 that the ratio between the curvature along the direction of negative curvature, d_k, and the smallest eigenvalue of H_k is uniformly bounded away from zero. Lemmas 5.5 and 5.6 imply that whenever d_k is nonzero, the ratio between the curvature along p_k and the smallest eigenvalue of H_k is also uniformly bounded away from zero.

In order to measure the magnitude of the curvature along d_k, it is compared to the smallest eigenvalue of H_k. Figure 8.1 shows the ratio between the curvature along d_k and the smallest eigenvalue of H_k for those iterates of the least-squares problems where d_k was nonzero. Figure 8.2 shows the corresponding data for the barrier problems. If 10% of the best possible curvature is regarded as "good", we see that this "good" curvature is computed in 95% of the cases for the least-squares problems and in 98% of the cases for the barrier problems.

Ideally, if d_k is nonzero, p_k should be both a nontrivial direction of negative curvature and a descent direction that is not too orthogonal to the negative gradient. Unfortunately, a direction that simultaneously has both these properties may not exist. In Figures 8.3 and 8.4 we give the ratio between the curvature along p_k and the curvature along d_k for iterates for which d_k was nonzero. Since d_k is intended to be a good direction of negative curvature, this ratio gives an idea of how much of the best possible curvature is achieved along p_k. The data for the least-squares problems is given in Figure 8.3; data for the barrier problems is given in Figure 8.4. Note that for both classes of problem, the ratio is close to one in most cases. Moreover, a ratio greater than one is possible if $\|p_k\| < \|d_k\|$. A ratio greater than 0.1 is achieved in 98% of the cases for the least-squares problems and in 99% of the cases for the barrier problems.

The direction s_k is intended to be a good descent direction. In order to investigate whether this property is inherited by p_k, the ratio of the cosine between s_k and g_k and the cosine between p_k and s_k was measured. These ratios are given in Figure 8.5 for the least-squares problems and in Figure 8.6 for the barrier problems. In general, this ratio is not as close to one as the curvature ratios. However, a ratio greater than 0.1 is obtained in 86% of the cases for the least-squares problems and in 93% of the cases for the barrier problems. We believe that the main reason for this ratio not being as close to one as the other two ratios, is that $\|s_k\|$ tends to zero as the solution is approached, but a nonzero $\|d_k\|$ will be of order one. Therefore, because of the way p_k is constructed, d_k will usually dominate s_k so that $p_k \approx d_k$.

It is noticeable that the barrier ratios seem better than the least-squares ratios. This is probably due to the fact that even though both problem classes contain highly nonlinear problems, the condition number of H_{11} is generally smaller for the barrier problems.

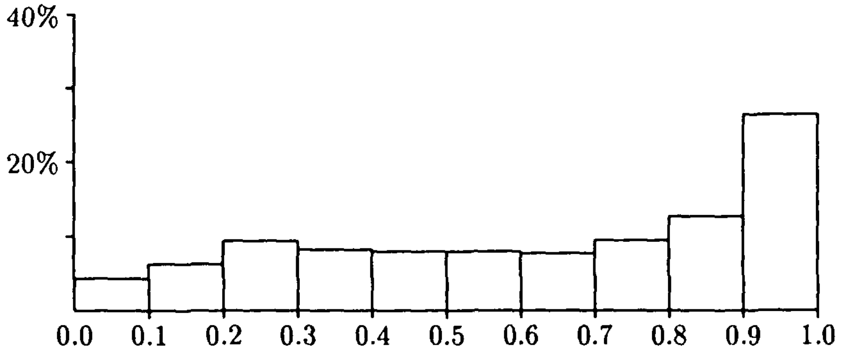

Figure 8.1: Least-squares problems: Ratio of the curvature along d_k to the smallest eigenvalue of the Hessian. Percentage out of 2013 observations.

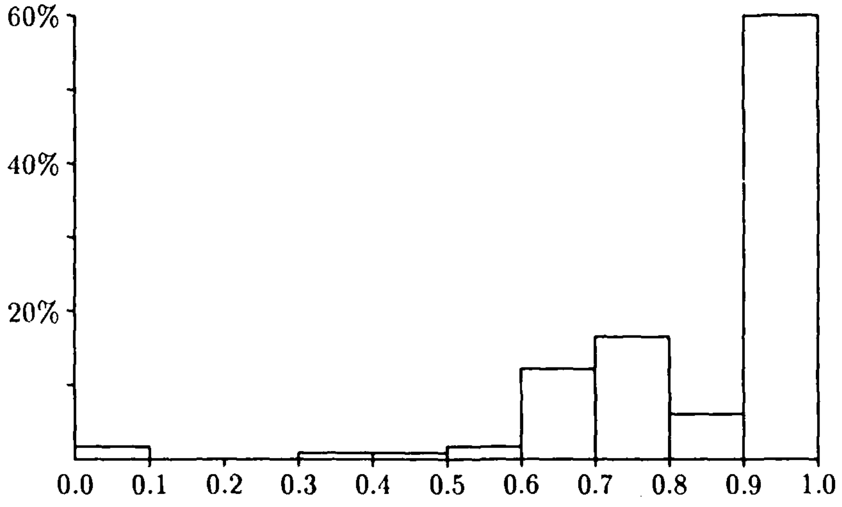

Figure 8.2: Barrier problems: Ratio of the curvature along d_k to the smallest eigenvalue of the Hessian. Percentage out of 115 observations.

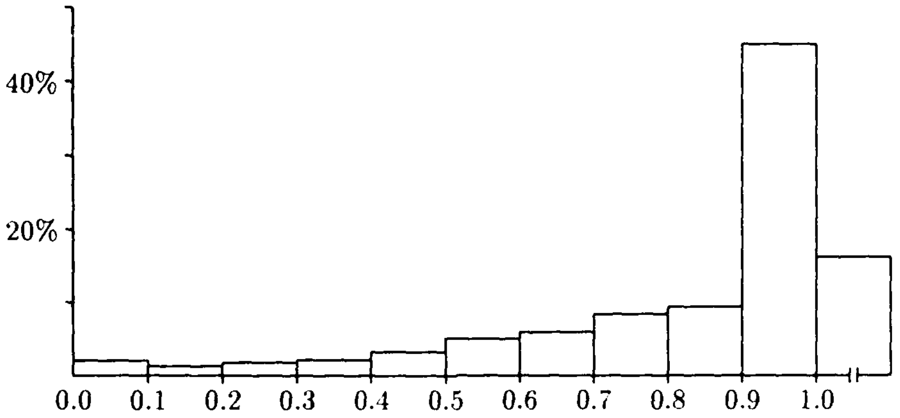

Figure 8.3: Least-squares problems: Ratio of the curvature along p_k to the curvature along d_k. Percentage out of 2013 observations.

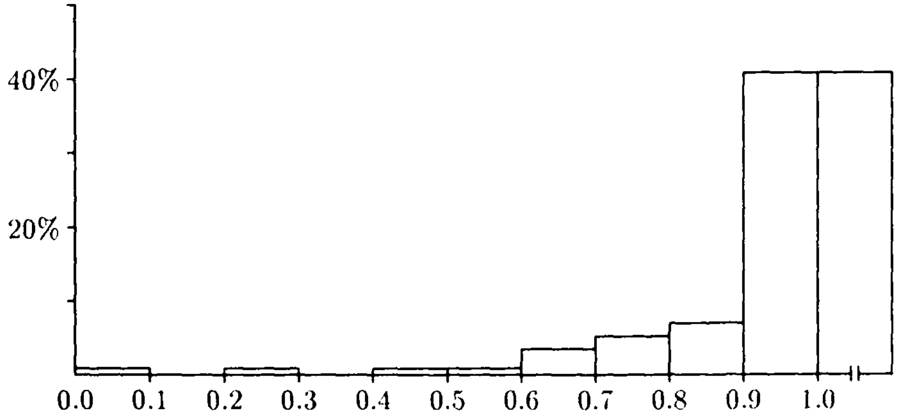

Figure 8.4: Barrier problems: Ratio of the curvature along p_k to the curvature along d_k. Percentage out of 115 observations.

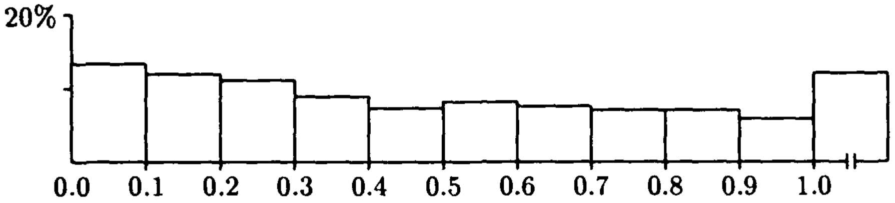

Figure 8.5: Least-squares problems: Ratio of the cosine between p_k and g_k to the cosine between s_k and g_k. Percentage out of 2013 observations.

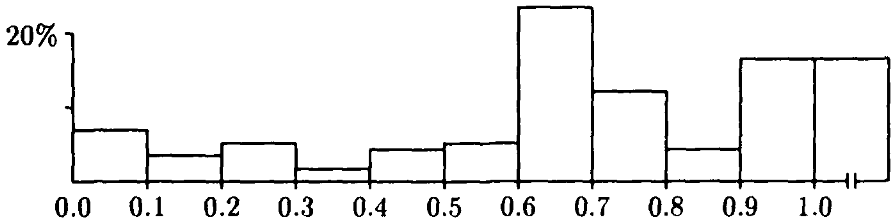

Figure 8.6: Barrier problems: Ratio of the cosine between p_k and g_k to the cosine between s_k and g_k. Percentage out of 115 observations.

9. Discussion

This report describes a modified Newton method for unconstrained minimization. At each iteration a positive-definite portion of the Hessian is factorized using the Cholesky algorithm. A descent direction is computed if the gradient is nonzero, and a direction of negative curvature is computed if the Hessian is sufficiently indefinite. A linear combination of these vectors define a search direction, along which the next iterate is found. Theoretical properties of the algorithm are established, and numerical data from a set of test problems are included.

As the algorithm is stated, if the direction of negative curvature is nonzero, it is always used to form p_k. From a practical point of view it is not clear if this is the best use of d_k. It is possible to define a number G such that whenever $\|g_k\| > G$, we may discard the direction of negative curvature and still obtain the convergence properties of Section 7. This alternative algorithm would allow a scheme for controlling that the cosine between p_k and g_k is not much smaller than the cosine between s_k and g_k, hereby ensuring that the search direction is not significantly closer to orthogonality to the negative gradient than the descent direction. The significance of utilizing a direction of negative curvature was investigated by rerunning the test problems with β_k set to zero for all k. On those problems where a direction of negative curvature previously had been used, the number of iterations required to satisfy the reduction of f_k given in Table 8.3 tended to increase. Moreover, the number of problems for which the convergence criteria were not met increased from six to twelve.

Finally, we note that the convergence results given in Section 7 imply convergence to a point where the gradient vector is zero and the Hessian matrix has a smallest eigenvalue greater than a small negative number. Since a point satisfying the second-order necessary conditions has nonnegative Hessian eigenvalues, it might appear that the convergence results are somewhat less satisfactory than those usually given for methods of this type. However, we observe that the magnitude of the bound on the smallest eigenvalue may be made as small as required by assigning a suitably small value for the parameter ϵ. Small values of ϵ affect only the numerical performance of the method, and not the theoretical convergence properties. Moreover, it may be observed from Tables 8.1, 8.2 and 8.4 that in most cases the iterates converged to a point where n_1 was equal to n, that is the Hessian was positive definite. The only exceptions are problems where the Hessian at the solution is very ill-conditioned, singular or undefined (for some of the barrier problems).

Our overall conclusion from the results is that it was possible to reduce the value of the objective function significantly in a rather small number of iterations by using directions of negative curvature whenever the Hessian was indefinite. However, to meet stringent convergence criteria was not always possible when the Hessian was very ill-conditioned or singular at the solution. Also, by running the algorithm without using the direction of negative curvature, we have the impression that the ability to compute a direction of negative curvature is not only a theoretical tool to show convergence, but also a helpful device in order to improve robustness and efficiency.

Acknowledgements

We would like to thank Christina Fraley for providing the Fortran subroutines for function evualuation of the least-squares problems. We also thank Per Olov Lindberg for his careful reading of the manuscript and a number of helpful comments.

A. Appendix: Properties of the Barrier Test Problems

The barrier test problems originate from a barrier function approach proposed by Resende *et al.* [RKR89] for solving 0–1 integer programs. Given an $m \times n$ matrix A and an n-vector b, the 0–1 integer program concerns finding a z in $\Re^n$ that belongs to the set F_1, where

$$F_1 = \{z : Az \leq b, \quad z_i = 0 \quad \text{or} \quad z_i = 1 \quad \text{for} \quad i = 1, \dots n\}.$$

We shall consider only the case where A and b have integer coefficients, since the analysis is simpler in this case.

Applying the linear transformation $x = 2z - e$, the 0–1 problem is transformed to an equivalent problem where all components of the integer solution are ±1. The transformed problem is converted into a smooth minimization problem by seeking values in the set F given by

$$F = \left\{ x : \begin{pmatrix} A \\ -I \\ I \end{pmatrix} x \leq \begin{pmatrix} 2b - Ae + e \\ e \\ e \end{pmatrix} \right\} = \{x : \bar{A}x \leq \bar{b}\}.$$

The aim is to find a point x^* in F with components ±1 by minimizing the barrier function

$$f(x) = \frac{1}{2} \ln(n - x^T x) - \frac{1}{m + 2n} \sum_{i=1}^{m+2n} \ln(e_i^T(\bar{b} - \bar{A}x))$$

(see Resende *et al.* [RKR89]).

Although the barrier function is not defined for a point x^* with components ±1, the following lemma shows that the barrier function has a global minimizer at x^*, since there exist sequences converging to x^* for which the function tends to minus infinity.

Lemma A.1. *Assume that the matrix A has at least one row. Furthermore, assume that $\{x_k\}_{k=0}^{\infty}$ converges to a point $x^* \in F$ such that x^* has all components ±1. If $\bar{A}x_k < \bar{b}$ for all k, and if there exists a positive constant c independent of k, such that it holds for all k that*

$$\min_i \frac{|e_i^T(x_k - x^*)|}{\|x_k - x^*\|} > c,$$

then $\lim_{k \to \infty} f(x_k) = -\infty$.

Proof. For clarity, the iteration subscript k is dropped when subscript i is used to denote a particular component of x_k.

Using properties of logarithms, $f(x)$ may be rewritten as

$$f(x) = \frac{1}{m+2n} \ln \left(\frac{(\sum_{i=1}^{n}(1 - x_i^2))^{m/2+n}}{\prod_{i=1}^{m} e_i^T(2b - Ae + e - Ax) \prod_{i=i}^{n}(1 - x_i^2)} \right). \tag{A.1}$$

Since $x^* \in F$ with all components ±1, it follows that $2b - Ae - Ax^* \geq 0$ and consequently

$$\lim_{k \to \infty} e_i^T(2b - Ae + e - Ax_k) \geq 1.$$

Therefore, it may be assumed without loss of generality that $Ax_k < 2b - Ae + e$ for all k. If r denotes the vector whose i-th component is given by $r_i = |x_i - x_i^*|$ we get

$$\frac{(\sum_{i=1}^{n}(1 - x_i^2))^n}{\prod_{i=1}^{n}(1 - x_i^2)} = \frac{(\sum_{i=1}^{n} r_i(2 - r_i))^n}{\prod_{i=1}^{n} r_i(2 - r_i)},$$

where it without loss of generality may be assumed that $r_i \leq 1$ for all i. Dividing both numerator and denominator by the positive quantity $(e^T r)^n$ and using the existence of the constant c, we derive the inequality

$$\frac{\left(\displaystyle\sum_{i=1}^{n} \frac{r_i}{e^T r}(2 - r_i) \right)^n}{\displaystyle\prod_{i=1}^{n} \frac{r_i}{e^T r}(2 - r_i)} < \left(\frac{2n^{3/2}}{c} \right)^n.$$

If $m > 0$ then $\lim_{k \to \infty}(n - x_k^T x_k)^{m/2} = 0$, and it follows that the argument of the logarithm in (A.1) tends to zero as k tends to infinity. Consequently, $\lim_{k \to \infty} f(x_k) = -\infty$ as required. ▮

The following lemma shows that there is a one-to-one correspondence between points in F with all components ±1 and points in F_1.

Lemma A.2. *The set F_1 is nonempty if and only if there exists a point in F with all components ±1.*

Proof. Assume that $z \in F_1$. A linear transformation $x = 2z - c$ yields $x \in F$ with all components ±1.

Assume that $x \in F$ with all components ±1. Let $x = 2z - e$, and it follows that z is a vector with all components zero or one, for which $Az \leq b + \frac{1}{2}e$. However, Az and b are integer vectors, and therefore it holds that $Az \leq b$. Consequently, $z \in F_1$. ▮

Lemma A.1 implies that the barrier function has a global minimizer at any point x^* in F with all components ±1. From Lemma A.2 it follows that if such a global minimizer is found, a point in F_1 may be identified, thereby providing a solution of the original 0–1 integer program.

References

[Arm66] L. Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16, 1-3, 1966.

[BK77] J. R. Bunch and L. Kaufman. Some stable methods for calculating inertia and solving symmetric linear systems. *Mathematics of Computation*, 31, 163-179, 1977.

[BP71] J. R. Bunch and B. N. Parlett. Direct methods for solving symmetric indefinite systems of linear equations. *SIAM J. on Numerical Analysis*, 8, 639-655, 1971.

[Cot74] R. W. Cottle. Manifestations of the Schur complement. *Linear Algebra and its Applications*, 8, 189-211, 1974.

[DGV85] J. E. Dennis, Jr., D. M. Gay, and P. A. Vu. *A new nonlinear equations test problem.* Technical Report 83-16, Department of Mathematical Sciences, Rice University, 1985.

[dVG81] N. de Villiers and D. Glasser. A continuation method for nonlinear regression. *SIAM J. on Numerical Analysis*, 18, 1139-1154, 1981.

[FF77] R. Fletcher and T. L. Freeman. A modified Newton method for minimization. *J. Optimization Theory and Applications*, 23, 357-372, 1977.

[FGM89] A. L. Forsgren, P. E. Gill, and W. Murray. *On the identification of local minimizers in inertia-controlling methods for quadratic programming.* Report SOL 89-11, Department of Operations Research, Stanford University, 1989.

[FM68] A. V. Fiacco and G. P. McCormick. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques.* John Wiley and Sons, Inc., New York, London, Sydney and Toronto, 1968.

[Fra88] C. Fraley. *Software performance on nonlinear least-squares problems.* Report SOL 88-17, Department of Operations Research, Stanford University, 1988.

[GM74] P. E. Gill and W. Murray. Newton-type methods for unconstrained and linearly constrained optimization. *Mathematical Programming*, 7, 311-350, 1974.

[GMSW79] P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright. *Two steplength algorithms for numerical optimization.* Report SOL 79-25, Department of Operations Research, Stanford University, 1979.

[GMW81] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization.* Academic Press, London and New York, 1981.

[Gol80] D. Goldfarb. Curvlinear path steplength algorithms for minimization which use directions of negative curvature. *Mathematical Programming*, 18, 31-40, 1980.

[GV83] G. H. Golub and C. F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, Baltimore, Maryland, 1983.

[Hig87] N. J. Higham. *Analysis of the Choleski decomposition of a semi-definite matrix.* Numerical Analysis Report 128, Department of Mathematics, University of Manchester, England, 1987.

[KD79] S. Kaniel and A. Dax. A modified Newton's method for unconstrained minimization. *SIAM J. on Numerical Analysis*, 16, 324-331, 1979.

[McC77] G. P. McCormick. A modification of Armijo's step-size rule for negative curvature. *Mathematical Programming*, 13, 111-115, 1977.

[McK75] J. J. McKeown. Specialized versus general-purpose algorithms for minimising functions that are sums of squared terms. *Mathematical Programming*, 9, 57-68, 1975.

[MGH81] J. J. Moré, B. S. Garbow, and K. E. Hillstrom. Testing unconstrained optimization software. *ACM Transactions on Mathematical Software*, 7, 17-41, 1981.

[MP78] H. Mukai and E. Polak. A second-order method for unconstrained optimization. *J. Optimization Theory and Applications*, 26, 501-513, 1978.

[MS79] J. J. Moré and D. C. Sorensen. On the use of directions of negative curvature in a modified Newton method. *Mathematical Programming*, 16, 1–20, 1979.

[OR70] J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York, 1970.

[RKR89] M. G. C. Resende, N. K. Karmarkar, and K. G. Ramakrishnan. An interior point algorithm for 0-1 integer programming. April 3-5 1989. Presented at the Third SIAM Conference on Optimization, Boston, Massachusetts.

[Sal87] D. E. Salane. A continuation approach for solving large residual nonlinear least squares problems. *SIAM J. on Scientific and Statistical Computing*, 8, 655–671, 1987.

[SE88] R. B. Schnabel and E. Eskow. *A new modified Cholesky factorization*. Technical Report CU-CS-415-88, Department of Computer Science, University of Colorado, Boulder, 1988.

[Wil65] J. H. Wilkinson. *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford, 1965.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
SOL 89-12

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
A Modified Newton Method for
Unconstrained Minimization

5. TYPE OF REPORT & PERIOD COVERED
Technical Report

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Anders L. Forsgren, Philip E. Gill and
Walter Murray

8. CONTRACT OR GRANT NUMBER(s)
N00014-87-K-0142

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Operations Research - SOL
Stanford University
Stanford, CA 94305-4022

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research - Dept. of the Navy
800 N. Quincy Street
Arlington, VA 22217

12. REPORT DATE
July 1989

13. NUMBER OF PAGES
34 pages

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale; its distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Unconstrained minimization, modified Newton method, negative curvature, Cholesky factorization, linesearch, steplength algorithm.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(see reverse side)

SOL 89-12: A Modified Newton Method for Unconstrained Minimization, Anders L. Forsgren, Philip E. Gill, and Walter Murray (July 1989, 34 pp.).

Newton's method has proved to be a very efficient method for solving strictly convex unconstrained minimization problems. For the nonconvex case, various *modified* Newton methods have been proposed

In this paper, a new modified Newton method is presented. The method is a linesearch method, utilizing the Cholesky factorization of a positive-definite portion of the Hessian matrix. The search direction is defined as a linear combination of a descent direction and a direction of negative curvature Theoretical properties of the method are established and its behaviour is studied when applied to a set of test problems.